J Exp Criminol (2012) 8:415–431
DOI 10.1007/s11292-012-9153-0

Can "disciplined passion" overcome the cynical view? An empirical inquiry of evaluator influence on police crime prevention program outcomes

**Brandon C. Welsh · Anthony A. Braga ·
Meghan E. Hollis-Peel**

Published online: 12 May 2012

Abstract

Objectives Investigate the degree and nature of influence that researchers have in police crime prevention programs and whether a high degree of influence is associated with biased reporting of results.

Methods Meta-analytic inquiry of experimental and quasi-experimental studies (n=42), drawn from four Campbell Collaboration systematic reviews of leading police crime prevention strategies: problem-oriented policing, "hot spots" policing, "pulling levers" policing, and street-level drug enforcement.

Results Larger program effects are not associated with studies with higher involvement on the part of the evaluator (e.g., assisting in strategy design, monitoring implementation, overcoming implementation problems).

Conclusions This study does not find support for the cynical view, which holds that researchers have a personal stake in the program or are pressured to report positive results. Importantly, the evaluator's involvement in the implementation of the program may be a necessary condition of successfully executed police experiments in complex field settings.

B. C. Welsh · M. E. Hollis-Peel
Northeastern University, Boston, MA, USA

B. C. Welsh · M. E. Hollis-Peel
Netherlands Institute for the Study of Crime and Law Enforcement, Amsterdam, The Netherlands

A. A. Braga
Rutgers University, Newark, NJ, USA

A. A. Braga
Harvard University, Cambridge, MA, USA

B. C. Welsh (✉)
School of Criminology and Criminal Justice, Northeastern University, Churchill Hall, 360 Huntington
Avenue, Boston, MA 02115, USA
e-mail: b.welsh@neu.edu

Keywords Academic–practitioner partnerships · Conflict of interest · Crime prevention · Evaluator influence · Policing

Rigorous program evaluation is a key component of prevention science, as scientific evidence helps policymakers and practitioners make well-informed decisions that ultimately improve people's lives. Evaluator objectivity is a core value in the creation of unbiased scientific evidence. Indeed, biased results lead to inappropriate decisions that, in turn, can produce harmful effects. Bias in program evaluation can be minimized through a variety of approaches such as adhering to rigorous scientific methods, ensuring broad participation of colleagues and others in study design, and avoiding conflicts of interests.

There are divergent views on how closely program evaluators, such as external academic researchers, should be involved with practitioners in program development and implementation. To some observers, close working relationships between practitioners and academics may violate the purported scientific necessity to separate program developer and evaluator roles (Eisner 2009a). To others, unless there is some convincing evidence of widespread evaluator bias or conflict of interest associated with such arrangements, these collaborations seem necessary to put academics in the position of being able to conduct high quality evaluations of prevention and intervention programs. As David Olds (2009) argues in his recent essay in support of "disciplined passion," balancing scientific integrity with the practical challenges associated with program evaluation in real world settings needs to be addressed through higher standards for reporting trials, better peer review, improved investigator training, and rigorous collegial support of those who choose this line of work.

Conflict of interest has been defined as a "set of conditions in which professional judgment concerning the validity of research might be influenced by a secondary competing interest" (Gorman and Conde 2007: 422). These secondary competing interests or the sources of conflict of interest are wide ranging and could include the prospects of financial gains, the advancement of ideological beliefs, or securing desired organizational changes. Conflict of interest is a profoundly important concern in research in the social and behavioral sciences as well as in the medical and physical sciences. Just the mention of it alone can raise questions about the results of a study and even the integrity of the researcher. The need for researchers to divest themselves of any conflict of interest is on par with our profession's oath "to do no harm." Indeed, any impropriety in the scientific process may end up causing harm to study participants.

Conflict of interest is serious business for sure, but it can also be complicated. Studies in the biomedical sciences consistently report a strong positive association between financial conflict of interest and research outcomes, with odds ratio effect sizes ranging from 3 to 5 (Eisner and Humphreys 2011). Pharmaceutical companies funding randomized trials of their own drugs are a commonplace example used to showcase this form of conflict of interest. While disclosure of financial conflict of interest is the most common means used to guard against bias in the scientific process, Allison (2009) reports that there is a growing movement of scientists and even a body of literature that questions this practice. Arguments against disclosure range from the politically charged: "judging scientists' credibility by their associations is tantamount to McCarthyism,” to the more introspective: “financial interests are neither the sole nor necessarily the most compelling motives” (Allison 2009: 522).

A crucial issue is how to establish if there is evidence of researcher bias in one form or another. Can bias be inferred when substantially larger effects are observed in studies conducted by intervention developers who serve as evaluators of their own intervention? This may be too simplistic. Instead of presuming bias, Lipsey (1995) put forward two plausible interpretations associated with this finding. One is the cynical view, which holds that these researchers have a personal stake in the program or are pressured to report positive results. The other is the high fidelity view, which holds that larger effects are a product of the researcher being able to attain a high degree of treatment integrity or fidelity to the model, a by-product of what Olds (2009) calls “disciplined passion.”

This article reports on an empirical inquiry of evaluator influence in a subset of leading criminological interventions. It is chiefly interested in investigating the degree and nature of influence that researchers have in police crime prevention programs and whether a high degree of influence is associated with biased reporting of results. Our sample is drawn from the highest quality evaluations of 4 innovative strategies of policing for crime prevention. These include problem-oriented policing (or POP), “hot spots” policing, “pulling levers” policing, and street-level drug enforcement. In POP, officers strive to use a basic iterative approach (problem identification, analysis, response, assessment, and adjustment of the response) to uncover the complex mechanisms at play in crime problems and to develop tailor-made interventions to address the underlying conditions that cause crime problems (Goldstein 1990). Hot spots policing challenges police officers to focus their resources and crime control efforts on the small number of specific places, such as clusters of addresses or streets and intersections, which generate a bulk of a city’s crime problem (Braga and Weisburd 2010). Pulling levers policing, also known as focused deterrence strategies, draws upon the POP framework and deterrence principles to focus criminal justice and social service attention on a small number of chronic offenders responsible for generating a disproportionate amount of a targeted crime problem such as gang-involved gun violence or disorderly street-level drug markets (Kennedy 2008). Street-level drug enforcement draws upon a diverse range of strategies such as enforcement crackdowns and community policing efforts to address street-level drug markets and their associated crime and disorder problems.

Our decision to investigate evaluator influence with this form of crime prevention rests largely on several key factors. One is that each one of these policing approaches is grounded in the action research model (Lewin 1946). Also, these policing approaches are considered among the most promising and effective in reducing disorder, property crime, and violent crime (Skogan and Frydl 2004). Yet another factor is that policing interventions have not been used in prior research on this topic (see below).

Literature review

Unlike in medicine and public health (see, e.g., Bekelman et al. 2003; Witt and Gostin 1994), there is a real paucity of published research on conflict of interest in criminology and criminal justice. The best known study was conducted by Geis et al. (1999). It was principally concerned with financial conflict of interest. As a case study, it focused on one criminology professor's involvement in an evaluation study of a private prison at the same time that he was receiving financial compensation to act as a consultant and advocate for the private prison industry.

Beyond its rather provocative style and the sharply worded rejoinder that soon followed (Lanza-Kaduce et al. 2000; see also Geis et al. 2000), the case study drew attention to a number of key points on the substantive topic. One was that there was a general lack of research on or concern with conflicts of interest in criminology and criminal justice at a point in time when there appeared to be an increasing number of possible incidents. It was also suggested that financial matters are the most prominent and worrisome element in instances of conflict of interest. The authors even went so far as to advance a guideline or principle on financial conflict of interest: "failure to come forward and disclose in a timely manner what later might be seen as bias based on financial self-interest inevitably will taint and call into question publications and statements that are not accompanied by such stipulations, however accurate the material and however pure the intent" (Geis et al. 1999, p. 375).

Several studies have been carried out more recently. Petrosino and Soydan (2005) and Eisner (2009a) investigated the potential impact of researcher involvement on program effects, with the following question in mind: Does an intervention produce a larger positive impact if it is evaluated by the same person or team that developed (and even implemented) it compared to an independent evaluation? Gorman and Conde (2007) also investigated the influence of developer-as-evaluator, but did not examine how this relates to program effects. As noted above, one interpretation of an affirmative finding to this question is that the researcher is biased in some fashion, what Lipsey (1995) called the cynical view. Another plausible interpretation is that the researcher helped to ensure a high level of treatment integrity.

Petrosino and Soydan (2005) carried out two sets of analyses to investigate the impact of program developers-as-evaluators on criminal recidivism outcomes. They first reviewed prior meta-analyses that looked at this issue. Of the 50 meta-analyses reviewed, only 12 tested the impact of evaluator influence on reported effect sizes, with 11 of these 12 meta-analyses reporting "larger effects for 'involved evaluators' in the treatment/program setting" (p. 438). The authors were careful to note that not one of the meta-analyses specifically examined program developers. Instead, the variable was categorized in a number of different ways, including "'influence of experimenter/investigator on treatment setting' or 'program evaluator independent of program'" (p. 438). The authors were satisfied with the close approximation of these variables to their variable of interest and the clear pattern across the 11 meta-analyses, concluding that "evaluators involved or influential in the program setting report larger effect sizes than evaluators who are not" (p. 439).

For their second set of analyses, the authors carried out their own meta-analysis, drawing upon a dataset of 300 randomized experiments on criminal justice. Interestingly, they excluded area- or placed-based experiments on policing and security. Three variables were analyzed: (1) the role of the evaluator as an internal or external researcher or collaboration between the two; (2) the influence of the evaluator on the intervention setting (high, moderate, or low); and (3) the specific role of the evaluator in the setting (e.g., program developer, external academician) (Petrosino and Soydan

2005: 441). In moving from analysis of variables 1 to 3, the authors sought to provide more specific information on the potential influence of developers-as-evaluators. Analyses of each of the three variables showed that studies with developers-as-evaluators produced substantially larger mean effect sizes than studies with independent or other types of evaluators. Petrosino and Soydan were careful to note that their data did not suggest an explanation for these findings, and instead discussed the findings in the context of Lipsey's (1995) cynical and high fidelity views of researcher influence.

Unlike Petrosino and Soydan (2005), but similar to Gorman and Conde (2007), Eisner (2009a) approached the subject with a high degree of professional skepticism. Perhaps coupled with his conclusion, this had the beneficial effect of an excellent scholarly exchange in the *Journal of Experimental Criminology*, with contributions from Olds (2009), Sherman and Strang (2009), and a final comment from Eisner (2009b). Eisner's (2009a) focus was more theoretical than empirical. He was also keen to advance "several strategies to examine empirically the extent of systematic bias related to conflict of interest" (p. 163). A more recent feasibility study by Eisner and Humphreys (2011) builds upon some of these strategies and proposes a coding instrument to assess potential financial conflict of interest, which could be built into criminological systematic reviews and meta-analyses. A preliminary test of their aggregate scale found some support for the view that financial conflict of interest is associated with larger program effects.

Eisner's (2009a) main conclusion was that there exists "circumstantial evidence suggesting that there might be a substantial problem" with conflict of interest in evaluations of criminological interventions (p. 165). This was based on his review of three separate data sources. One was the meta-analysis and review of meta-analyses by Petrosino and Soydan (2005). Another source involved a couple of methodological case studies of systematic reviews of interventions with crime outcomes. The review by Gorman and Conde (2007), which is discussed below, was included among these case studies. The third source involved comparisons between developer-led and independent evaluations of 4 well-known programs in the areas of indicated drug prevention, early parenting skills training, school-based bullying prevention, and substance abuse prevention.

Olds (2009), in his response, argued that innovation depends upon integrating the role of intervention developer and investigator, as innovation and commitment to follow-through depends upon overriding commitment to improving outcomes. Self-interest of developer may be offset by determination to solve problems, efforts to raise money for research, and knowledge of vulnerabilities of interventions. Sherman and Strang (2009) echoed this view and discussed how the perception of bias is not just relegated to the developer as investigator, but can also occur among "independent evaluators who may seek to 'get a scalp' of a developer or a program" (p. 185).

Gorman and Conde (2007) investigated the influence of developers-as-evaluators in the context of "model" or research-based programs on school-based drug and violence prevention. A two-stage process was used. Based upon a sample of 34 of these programs, the authors first examined the relationship that exists between program developers and the organizations that distribute the programs to schools. It was found that half the programs (17) exhibited the most direct form of a financial relationship, whereby the program developer “owns or directs the company that distributes the program (or provides training in it) or receives remuneration from a third party (typically a publishing company) that sells the program” (Gorman and Conde 2007: 426). In the next stage, the authors assessed all of the published evaluation reports related to these 34 programs, a total of 246 reports. It was found that 78 % (193 of 246) of the reports included the program developer as an author. While acknowledging that their study was “exploratory in nature,” Gorman and Conde expressed concern about the apparent lack of “complete separation between the program developer and program distributor” (p. 427), something they argue is needed in school-based prevention programs.

Methodology

Sample

Four Campbell Collaboration systematic reviews of the different policing interventions (POP, hot spots, pulling levers, and street-level drug enforcement) served as the basis for the sample of studies used here. Three of these reviews are published in Campbell’s electronic library (http://www.campbellcollaboration.org/reviews_crime_justice/index.php; Braga 2007; Mazerolle et al. 2007; Weisburd et al. 2008) and the other (on pulling levers policing) is in the final stage before publication (see Braga and Weisburd 2011). Braga’s (2007) hot spots policing review has been updated and is also in the final stage before publication; new studies were available to us.

From these systematic reviews, only the highest quality evaluation studies were selected. This included randomized controlled trials and quasi-experimental designs,[1] incorporating before-and-after measures of crime in experimental and control conditions. Control conditions are needed to counter threats to internal validity (Shadish et al. 2002).

Coding instrument and protocol

A coding instrument was developed to assess the nature and degree of researcher involvement in the included studies. A number of general items (e.g.,

[1] The Campbell Collaboration reviews used to identify police crime prevention evaluations include strong and weak quasi-experimental designs. Based on the Maryland Scientific Methods Scale (Farrington et al. 2006), eligible quasi-experimental evaluations of police crime prevention programs would be considered “Level 3” and “Level 4” research designs. Level 3 quasi-experimental designs are regarded as the minimum design that is adequate for drawing conclusions about program effectiveness. This design rules out many threats to internal validity such as history, maturation/trends, instrumentation, testing, and mortality. The main problems of Level 3 evaluations center on selection effects and regression to the mean due to the non-equivalence of treatment and control conditions. Level 4 evaluations measure outcomes before and after the program in multiple treatment and control condition units. These types of designs have better statistical control of extraneous influences on the outcome and, relative to lower level evaluations, deal with selection and regression threats more adequately.

authors, date, published or not) and key study characteristics (e.g., research design) were also collected from the studies. Effect size statistics were drawn from the systematic reviews or calculated by the present authors. For each study calculation of the effect size and its variance adhered to standard conventions (Lipsey and Wilson 2001).

The coding instrument drew upon the 3 items used in Petrosino and Soydan (2005). Five additional items that are more specific to researcher involvement in police interventions served as the main component of the instrument. (The full instrument is available from the authors.) The first of these 5 items asked if the evaluator was involved in a prior training of the police officers who executed the program. The second and third items asked if the evaluator was involved in problem diagnosis and program/strategy design, respectively. The fourth item asked if the evaluator was involved in monitoring the implementation of the program, and the final one asked if the evaluator provided assistance in overcoming implementation problems. All of these items were coded as no, yes, or unknown. Only 1 unknown response was recorded.

A coding protocol was established by the research team. The first step involved the research team meeting to discuss in detail the coding instrument. (All 3 members have extensive experience in conducting systematic reviews.) Next, studies were randomly allocated among the 3 members of the team and studies were retrieved and coded by the individual members. The research team then met to discuss the coding of all of the studies and resolve any questions. In the event of missing information or the need for clarification about the coding, study authors were contacted by members of the research team.

Evaluator involvement scale

Following the procedures recommended by Spector (1992), the main 5 items in the coding instrument were used to develop a 6-point (0–5) summated rating scale for evaluator involvement, with 0 = no involvement and 5 = intense involvement. This allowed for an aggregate rating of lower evaluator involvement in program implementation (0–2) or higher evaluator involvement in program implementation (3–5).

Meta-analysis

Meta-analytic techniques were used to determine the size, direction, and statistical significance of the overall impact of the policing strategies on crime. Program effect sizes were weighted on the variance of the effect size and the study sample size (Lipsey and Wilson 2001). We used the standardized mean difference effect size, also known as Cohen's *d* (Cohen 1988). For each study, we used either a reported summary outcome (e.g., total crime incidents, total calls for service) or a solitary reported outcome if the evaluation only examined the effects of the police intervention on a single crime outcome (e.g., gun assaults, violent crime). If no summary outcome was available and multiple outcomes were reported, we combined all reported outcomes into an overall average effect size statistic. Biostat's Comprehensive Meta Analysis Version 2.2 was used.

Results

Across the 4 Campbell reviews of police crime prevention program evaluations, there were 46 unique studies.[2] Unfortunately, 4 studies could not be included in our meta-analysis: 2 did not provide the necessary details to calculate effect sizes (Criminal Justice Commission 1998; Hope 1994); 1 did not use appropriate statistical tests to evaluate program impacts[3] (Caeti 1999); and the other, a pulling-levers focused deterrence evaluation, did not test a police-involved program (Hawken and Kleiman 2009).

Of the 42 included studies, 33 (78.6 %) were quasi-experimental evaluations and 9 (21.4 %) were randomized controlled trials. The evaluator connection to the study police department was characterized as external evaluator only in 34 (81.0 %) of the studies, collaboration of external and internal police evaluators in 6 (14.3 %) of the studies, and internal police evaluator only in 2 (4.8 %) of the studies. Table 1 presents the frequencies of the items that comprised the evaluator involvement scale. The least frequency category of researcher activity in the eligible studies was evaluator involvement in prior training of officers who executed the program (31.0 %). Evaluator involvement in problem diagnosis, program/strategy design, monitoring program implementation, and working with practitioners to overcome program implementation difficulties occurred in similar frequencies. The evaluator involvement scale had a mean = 1.5 and exhibited a bimodal distribution with the 0 category having 16 studies (or 38.1 %) and the 5 category having 12 studies (or 28.6 %). The evaluator involvement scale had a Cronbach's Alpha = .833, suggesting good internal consistency across the items that comprised the scale.

Overall effects of police crime prevention programs

Using the mean effect criterion for the calculation of program effects of the 42 eligible studies, the forest plots in Fig. 1 show the standardized difference in means between the treatment and control or comparison conditions (effect size) with a 95 % confidence interval plotted around them for all tests. Points plotted to the right of 0

[2] Nine of these 46 studies appeared in multiple systematic reviews. This was not surprising given that these 4 systematic reviews were related in the sense that all examined evaluations of innovative police crime prevention programs focused on specific crime problems. The Mazerolle et al. (2000) and Weisburd and Green (1995) evaluations were included in the problem-oriented policing, drug enforcement, and hot spots policing systematic reviews. The Sherman and Rogan (1995b) and Sviridoff et al. (1992) evaluations were included in the drug enforcement and hot spots policing systematic reviews. The Braga et al. (1999) and Sherman et al. (1989) studies were included in the problem-oriented policing and hot spots policing systematic reviews. The Green (1996) and Clarke and Bichler-Robertson (1998) studies appeared in the problem-oriented policing and drug enforcement systematic reviews. The Braga et al. (2001) evaluation was included in the problem-oriented policing and pulling levers systematic reviews.

[3] The Caeti (1999) evaluation did not examine the differences-in-differences between treatment and control areas and, as such, did not directly measure whether the observed changes in the treatment beats were significantly different from observed changes in the control beats.

Table 1 Frequency distribution of evaluator activities in police crime prevention program studies (n=42)

Activity	n	%
Involved in training of police officers executing program	13	31.0
Involved in problem diagnosis	20	47.6
Involved in program/strategy design	21	50.0
Involved in monitoring program implementation	20	47.6
Involved in overcoming program implementation difficulties	21	50.0

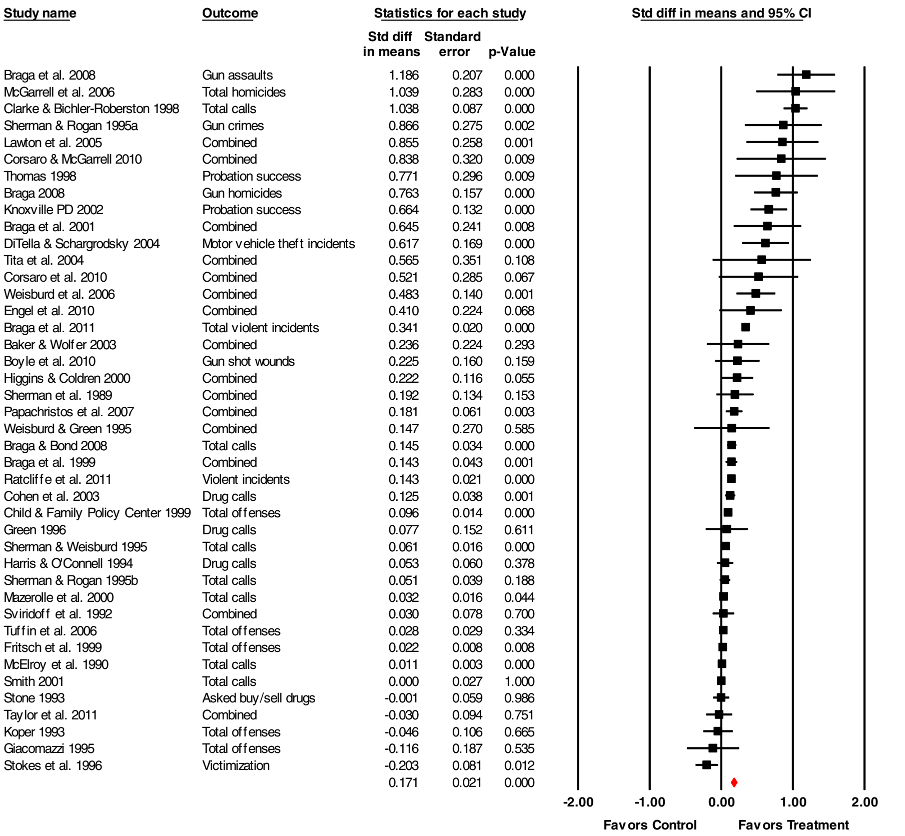
indicate a treatment effect; in this case, the test showed a reduction in crime or disorder. Points to the left of 0 indicate a backfire effect, whereby control conditions improved relative to treatment conditions. Since the Q statistic was statistically significant (Q=678.859, df=41, p<0.000), we used a random effects model to estimate the overall mean effect size based on a heterogeneous distribution of effect sizes. The meta-analysis of effect sizes suggests a highly significant effect in favor of these police crime prevention strategies (p<.001), with an overall effect size of .171.[4]

Publication bias

Publication bias presents a strong challenge to any review of evaluation studies (Rothstein 2008). Campbell reviews, such as the 4 used here, take a number of steps to reduce publication bias, as represented by the fact that 13 of the 42 studies in our meta-analysis came from unpublished sources. Wilson (2009) has argued that there is often little difference in methodological quality between published and unpublished studies, which points to the importance of searching the “grey literature.” We first compared the overall mean effect size of the 13 unpublished studies (.145, SE = .048, p<.05) to the overall mean effect size of the 29 published studies (.161, SE = .021, p<.05). The effect sizes were quite similar, suggesting that publication bias does not present a problem for our analyses.

We then used the trim-and-fill procedure (Duval and Tweedie 2000) to estimate the effect of potential data censoring, such as publication bias, on the outcome of the meta-analyses. The diagnostic funnel plot is based on the idea that, in the absence of bias, the plot of study effect sizes should be symmetric about the mean effect size. If there is asymmetry, the trim-and-fill procedure imputes the missing studies, adds them to the analysis, and then re-computes the mean effect size. A visual inspection of the funnel plot indicates some asymmetry with more studies with a large effect to the right of the mean than the left of the mean (see Fig. 2). The trim-and-fill procedure determined that 12 studies should be added to create symmetry. Using a random effects model, the mean effect size

[4] Using the guidelines on effect size interpretation suggested by Cohen (1988), a standardized mean effect size of .171 would be considered small. Other scholars, however, suggest a more nuanced interpretation of the magnitude of effect sizes in different social science fields. For instance, Lipsey (2000) suggests that a small standardized mean effect size should be defined as .10 rather than .20. Using Lipsey’s guidelines, the standardized mean effect size for the police crime prevention programs in this review would be considered moderate.



Study name	Outcome	Statistics for each study Std diff in means	Standard error	p-Value
Braga et al. 2008	Gun assaults	1.186	0.207	0.000
McGarrell et al. 2006	Total homicides	1.039	0.283	0.000
Clarke & Bichler-Roberston 1998	Total calls	1.038	0.087	0.000
Sherman & Rogan 1995a	Gun crimes	0.866	0.275	0.002
Lawton et al. 2005	Combined	0.855	0.258	0.001
Corsaro & McGarrell 2010	Combined	0.838	0.320	0.009
Thomas 1998	Probation success	0.771	0.296	0.009
Braga 2008	Gun homicides	0.763	0.157	0.000
Knoxville PD 2002	Probation success	0.664	0.132	0.000
Braga et al. 2001	Combined	0.645	0.241	0.008
DiTella & Schargrodsky 2004	Motor vehicle theft incidents	0.617	0.169	0.000
Tita et al. 2004	Combined	0.565	0.351	0.108
Corsaro et al. 2010	Combined	0.521	0.285	0.067
Weisburd et al. 2006	Combined	0.483	0.140	0.001
Engel et al. 2010	Combined	0.410	0.224	0.068
Braga et al. 2011	Total violent incidents	0.341	0.020	0.000
Baker & Wolfer 2003	Combined	0.236	0.224	0.293
Boyle et al. 2010	Gun shot wounds	0.225	0.160	0.159
Higgins & Coldren 2000	Combined	0.222	0.116	0.055
Sherman et al. 1989	Combined	0.192	0.134	0.153
Papachristos et al. 2007	Combined	0.181	0.061	0.003
Weisburd & Green 1995	Combined	0.147	0.270	0.585
Braga & Bond 2008	Total calls	0.145	0.034	0.000
Braga et al. 1999	Combined	0.143	0.043	0.001
Ratcliffe et al. 2011	Violent incidents	0.143	0.021	0.000
Cohen et al. 2003	Drug calls	0.125	0.038	0.001
Child & Family Policy Center 1999	Total offenses	0.096	0.014	0.000
Green 1996	Drug calls	0.077	0.152	0.611
Sherman & Weisburd 1995	Total calls	0.061	0.016	0.000
Harris & O'Connell 1994	Drug calls	0.053	0.060	0.378
Sherman & Rogan 1995b	Total calls	0.051	0.039	0.188
Mazerolle et al. 2000	Total calls	0.032	0.016	0.044
Sviridoff et al. 1992	Combined	0.030	0.078	0.700
Tuffin et al. 2006	Total offenses	0.028	0.029	0.334
Fritsch et al. 1999	Total offenses	0.022	0.008	0.008
McElroy et al. 1990	Total calls	0.011	0.003	0.000
Smith 2001	Total calls	0.000	0.027	1.000
Stone 1993	Asked buy/sell drugs	-0.001	0.059	0.986
Taylor et al. 2011	Combined	-0.030	0.094	0.751
Koper 1993	Total offenses	-0.046	0.106	0.665
Giacomazzi 1995	Total offenses	-0.116	0.187	0.535
Stokes et al. 1996	Victimization	-0.203	0.081	0.012
		0.171	0.021	0.000

Fig. 1 Meta-analysis of police prevention program effects on crime
Note: All references available in Campbell systematic reviews

decreased from 0.171 (95 % CI=0.130, 0.212) to 0.089 (95 % CI=0.047, 0.132) with the imputed studies. Despite the reduced mean effect size, it was still in the positive direction and statistically significant.[5]

[5] The slight overlap of the 95 % confidence intervals raised the possibility that the two parameters may in fact be different. As suggested by Rothstein (2008), we used a variety of tests to examine the possibility that our study suffered from some degree of publication bias. Our overall conclusion from these analyses was that publication bias was not a problem for our study. For instance, the classic failsafe *N* test yielded a *z* value of 18.70 and a corresponding *p* value of<0.000 for the combined test of significance. The test reported that there would need to be 3,782 missing studies with zero effect to yield a combined two-tailed *p* value exceeding 0.05. This far exceeds the 220 studies suggested by Rosenthal's (5 K+10) guideline on the number of studies to be confident that the results would not be nullified. We also applied the Begg and Mazumdar rank correlation test to our pool of studies; the Kendell's tau b for the 42 studies was 0.111 with a two-tailed *p*=0.298 (based on continuity corrected normal approximation), suggesting publication bias was not operating in the analyses. As a result of these supplementary analyses, we are confident that publication bias is not a significant problem for our study of evaluator influence on program outcomes.

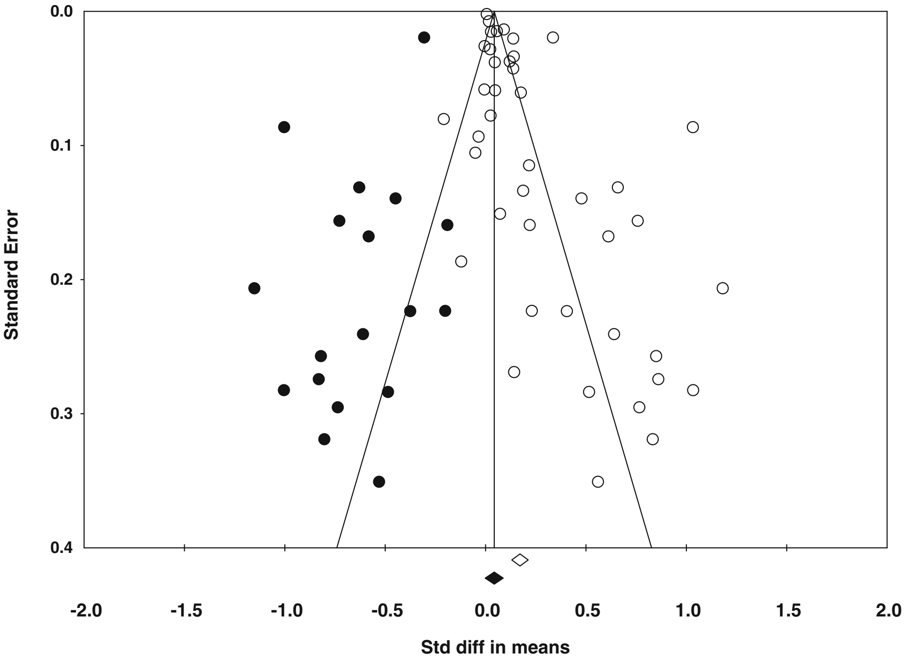


Fig. 2 Funnel plot of 42 included studies with imputed studies from trim-and-fill analysis
Note: *Empty circles* are the 42 included studies. *Filled-in circles* indicate inputed studies from trim-and-fill analysis

Moderator analyses

Moderator variables help to explain and understand differences across studies in the outcomes observed. Given the important distinction in methodological quality between the randomized controlled trials and quasi-experimental evaluation studies, we first examined research design as a moderator variable. Table 2 presents a random effects model examining the two different classes of evaluation designs included in this review. Consistent with prior research suggesting that weaker designs are more likely to report stronger effects in crime and justice studies (Weisburd et al. 2001; Welsh et al. 2011), the quasi-experimental designs were associated with a much larger within-group effect size (.221, p<.05) relative to the randomized controlled designs (.089, p<.05). This does not mean that quasi-experimental studies cannot be of high quality. Rather, it suggests that quasi-experimental designs in police crime prevention program evaluations are likely to overstate outcomes as contrasted with randomized experiments.

We then used random effects models to examine whether the connection of the evaluator to the police department implementing the program yielded differential impacts on overall mean effect sizes. Since there were only 2 studies that were completed by an internal police evaluator where an effect size could be calculated, we compared external evaluator only studies (n=34) with studies in which the evaluator was internal to the police department or an external evaluator worked directly with an internal police evaluator (n=8). As Table 2 reveals, the 95 % confidence intervals overlap for these two distinct categories of connections of the police department to the evaluation. This suggests that the mean effect sizes for the

Table 2 Moderator analyses of police prevention program outcomes (n=42)

	n	Effect size	95 % CI
	Design		
Randomized controlled trial	9	.089*	.048, .130
Quasi-experimental evaluation	33	.221*	.167, .274
	Evaluator connection to police department		
External evaluator	34	.176*	.132, .220
Internal/collaboration	8	.229*	.044, .415
	Level of implementation involvement		
Higher level of involvement (3–5 score)	20	.172*	.098, .246
Lower level of involvement (0–2 score)	22	.206*	.142, .270

* p<.05

categories may not be dissimilar. However, the internal/collaboration category was associated with a higher point estimate of the mean effect size (.229 vs. .176). This suggests that external researchers may conduct studies that provide slightly more conservative estimates of program impacts compared to studies influenced by internal police staff.

We next collapsed our 6-point evaluator involvement scale into two categories: lower evaluator involvement in program implementation (0–2, n=20 studies) and higher evaluator involvement in program implementation (3–5, n=22 studies). Table 2 shows the 95 % confidence intervals overlap for these two categories of evaluator involvement in the program being tested. This suggests that the mean effect sizes for the categories may actually be the same. However, it is also interesting to note that the point estimate for mean program effect size for evaluations with higher degrees of evaluator involvement (.172) is lower than the point estimate for evaluations with lower degrees of evaluator involvement (.206).

Eight of the 9 randomized controlled trials were completed by external evaluators only and 1, the Kansas City Crack House Raid Experiment (Sherman and Rogan 1995b), was completed by a team of internal and external evaluators. Clearly, police departments rely on external evaluators to conduct the most rigorous tests of their crime prevention programs. Relative to the quasi-experimental evaluations, the randomized controlled trials were characterized by higher levels of evaluator involvement in the implementation of police crime prevention programs. Table 3 reveals that almost half (45.5 %; n=15) of the quasi-experimental evaluations had no evaluator involvement and almost two-thirds (63.6 %; n=21) of the quasi-experiments had lower levels of evaluator involvement (0–2) in the implementation of the program. In contrast, more than half (55.6 %; n=5) of the randomized controlled trials had the highest level of evaluator involvement (5), and with the exception of 1 of the randomized experiments all had the highest level of evaluator involvement in the implementation of the program. Given the lower mean effect sizes associated with randomized experiments, these findings suggest that high degrees of evaluator involvement do not necessarily translate into inflated program effect size estimates. Instead, the evaluator’s involvement in the implementation of the program may be a

Table 3 Research design by evaluator involvement index (n=42)

Evaluator involvement index	Quasi-experiments	Randomized controlled trials
0	15 (45.5 %)	1 (11.1 %)
1	2 (6.1 %)	0 (0.0 %)
2	4 (12.1 %)	0 (0.0 %)
3	1 (3.0 %)	1 (11.1 %)
4	4 (12.1 %)	2 (22.2 %)
5	7 (21.2 %)	5 (55.6 %)
Total	33 (100 %)	9 (100 %)

necessary condition of successfully executed police experiments in complex field settings.

It is important to note here that this table suggests a relationship between research design and evaluator involvement in the implementation of programs. Since both are related to program effect size, these moderator variables are potentially confounded. Future inquiries should more closely examine evaluator involvement variables by adjusting for design type in a meta-analytic regression model. Ideally, this analysis would be expanded beyond police crime prevention programs to include a much larger number of crime and justice program evaluations. A regression analysis with a larger number of studies would be able to control for potentially biasing effects of research design and allow for a more precise estimate of the relationship between evaluator involvement and program effect size.

Discussion and conclusions

The production of objective scientific evidence to guide prevention policy and practice requires program evaluators to minimize all potential sources of bias that could negatively affect study findings. One potential source of bias may be generated by conflicts of interests that arise when evaluators become too invested in the program they are evaluating and lose scientific objectivity. We believe that our study advances the debate on whether developers-as-evaluators generate inflated estimates of program effects, by taking a more differentiated look at the specific activities of researchers involved in the evaluation of police crime prevention programs. In this article, we examined whether police crime prevention studies with more intensive evaluator involvement in program development and implementation produced larger program effects when compared to police crime prevention studies with low levels of evaluator involvement. Our findings suggest that more intensive evaluator involvement in program activities do not necessarily translate into inflated estimates of program effects. In fact, higher levels of evaluator involvement may be associated with more conservative estimates of program effects through the more rigorous research designs implemented in these closer researcher-practitioner partnerships.

Our analyses do suggest that evaluations that are executed by internal police staff or research teams that include internal police staff seem to generate slightly higher program effect estimates relative to external research teams. However, as noted above, these differences are not statistically significant. Further study is needed to determine whether internal police staff experience undue pressure to report positive findings that fit with official police department positions on the value of the programs in question. For the small number of higher quality evaluations reviewed here, this does not seem to be a meaningful problem. Regardless, we believe it is more important to distinguish and better understand the particular activities engaged by evaluators when drawing conclusions about the influence of higher degrees of evaluator involvement on program outcomes.

The activities that comprise our evaluator involvement scale are steps that researchers should take to ensure a high fidelity test of prevention and intervention programs. The evaluator assisted with conducting problem analyses to document the underlying conditions that the program sought to change, designing the program to ensure that it was appropriately focused on the units of analysis and adhered to evaluation requirements, training officers to ensure they understood the program rationale and activities, monitoring program implementation to determine whether the treatment dosage was applied adequately, and working with program managers to address any implementation obstacles that arose during the study time period. We believe these activities represent the "disciplined passion" necessary to evaluate social programs in complex field settings. Without this deep level of involvement in the field, evaluation findings and the conclusions drawn from them may suffer from unknown implementation and measurement problems associated with low fidelity tests of programs.

It is important to reaffirm that we do not believe that any one moderator variable can explain the effects of the innovative policing strategies reviewed here, or be explained away for that matter. Lipsey (2003) explored the difficulties of investigating and interpreting moderator variables in meta-analyses. He argued that, because moderator variables are generally related to each other and to the effect of the intervention (or effect size), it can be difficult to determine the influence of a single moderator on effect size. Lipsey referred to this as the confounding effect of moderator variables. Our efforts to investigate the influence of research design and publication bias went some toward this approach. Future research on evaluator influence should be mindful of this and examine other important moderator variables. Clearly, the relationship between the level of evaluator involvement and research design type needs to be further scrutinized.

It is also important to reiterate that the present study was interested in investigating the degree and nature of influence that researchers have in police crime prevention programs. Following Petrosino and Soydan (2005) and Eisner (2009a), it was not focused specifically on financial conflict of interest. Unlike other areas of criminal justice or social policy, none of these policing interventions are packaged programs commercially available to police departments. Thus, the nature of the potential conflicts of interest is different and may be somewhat less in this research domain.

The evaluator actions described above are part of the well established action research model that undergirds many criminal justice and public health responses to recurring public safety and health problems (Moore et al. 1994). As described by Reason and Bradbury (2008), action research is an interactive inquiry process that balances problem-solving actions implemented in a collaborative context with datadriven analysis or research to understand underlying causes that enable programmatic decisions about personal and organizational change. Just as decisions of doctors are supposed to be based on medical science, police should apply analytical approaches and interventions based on sound theory and evidence. The high-researcher-involvement police crime prevention evaluations reviewed here follow this well-grounded applied science approach.

Unfortunately, the overall state of crime and justice research is generally methodologically weak (National Research Council 2008). A growing number of scholars suggest that more opportunities could be created to conduct rigorous evaluations of crime and justice interventions by developing strong academic-practitioner research collaborations (Braga 2010; Petersilia 2008). Indeed, the most rigorous tests of police crime prevention programs—randomized controlled trials—were usually implemented with the considerable involvement of external academic researchers. Research partnerships allow academics to get their feet in the door, develop trust with practitioners, and position themselves to make a stronger argument for using rigorous evaluation designs such as randomized controlled trials. Petersilia (2008) suggests that policymakers and practitioners today are often willing to support true randomized experiments and are more likely to be influenced by experimental findings than in the past. Many higher-level managers have had research methods courses and most understand and are familiar with medical trials where new drugs are routinely tested with experimental designs (Petersilia 2008).

While our findings are limited to a group of police crime prevention program evaluations, we believe the broader lesson is that disciplined passion in conducting field evaluations can overcome the cynical view. The researcher activities that characterize more intensive involvement with evaluated programs seem necessary to ensure strong program tests in uncertain field conditions that do not inflate positive program findings. As such, without some definitive evidence of improprieties by the study investigator, results generated by evaluations flowing from these action research partnerships should be judged on the merits of the study design and execution rather than on the notion that results are biased because of the investigator's involvement in program design and implementation.

Acknowledgments We are grateful to the editor, David Wilson, and the anonymous reviewers for helpful comments.

References

Allison, D. B. (2009). The antidote to bias in research. *Science, 326*, 522–523.

Bekelman, J. E., Li, Y., & Gross, C. P. (2003). Scope and impact of financial conflicts of interest in biomedical research. *Journal of the American Medical Association, 289*, 454–465.

Braga, A. A. (2007). The effects of hot spots policing on crime. *Campbell Collaboration*. doi:10.4073/csr.2007.1.

Braga, A. A. (2010). Setting a higher standard for the evaluation of problem-oriented policing initiatives. *Criminology and Public Policy, 9*, 173–182.

Braga, A. A., & Weisburd, D. (2010). *Policing problem places: Crime hot spots and effective prevention*. New York: Oxford University Press.

Braga, A. A., & Weisburd, D. (2011). The effects of focused deterrence strategies on crime: A systematic review and meta-analysis of the empirical evidence. *Journal of Research in Crime and Delinquency*, 48, in press.

Braga, A. A., Weisburd, D., Waring, E., Mazerolle, L. G., Spelman, W., & Gajewski, F. (1999). Problem-oriented policing in violent crime places: a randomized controlled experiment. *Criminology, 37*, 541–580.

Braga, A. A., Kennedy, D. M., Waring, E. J., & Piehl, A. M. (2001). Problem-oriented policing, deterrence, and youth violence: an evaluation of Boston's Operation Ceasefire. *Journal of Research in Crime and Delinquency, 38*, 195–225.

Caeti, T. J. (1999). *Houston's targeted beat program*. Unpublished Ph.D. dissertation. Huntsville, TX: Sam Houston State University.

Clarke, R. V., & Bichler-Robertson, G. (1998). Place managers, slumlords and crime in low rent apartment buildings. *Security Journal, 11*, 11–19.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale: Erlbaum.

Criminal Justice Commission. (1998). *Beenleigh calls for service project*. Brisbane: Criminal Justice Commission.

Duval, S., & Tweedie, R. (2000). A nonparametric 'trim and fill' method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*, 89–98.

Eisner, M. (2009a). No effects in independent prevention trials: can we reject the cynical view? *Journal of Experimental Criminology, 5*, 163–184.

Eisner, M. (2009b). Reply to the comments by David Olds and Lawrence Sherman. *Journal of Experimental Criminology, 5*, 215–218.

Eisner, M., & Humphreys, D. (2011). Measuring conflict of interest in prevention and intervention research: A feasibility study. In T. Bliesener, A. Beelmann, & M. Stemmler (Eds.), *Antisocial behavior and crime: Contributions of developmental and evaluation research to prevention and intervention* (pp. 165–180). Cambridge: Hogrefe Publishing.

Farrington, D. P., Gottfredson, D. C., Sherman, L. W., & Welsh, B. C. (2006). The Maryland scientific methods scale. In L. W. Sherman, D. P. Farrington, B. C. Welsh, & D. L. MacKenzie (Eds.), *Evidence-based crime prevention* (rev. ed., pp. 13–21). New York: Routledge.

Geis, G., Mobley, A., & Shichor, D. (1999). Private prisons, criminological research, and conflict of interest: a case study. *Crime & Delinquency, 45*, 372–388.

Geis, G., Mobley, A., & Shichor, D. (2000). Letter to the editor. *Crime & Delinquency, 46*, 443–445.

Goldstein, H. (1990). *Problem-oriented policing*. Philadelphia: Temple University Press.

Gorman, D. M., & Conde, E. (2007). Conflict of interest in the evaluation and dissemination of 'model' school-based drug and violence prevention programs. *Evaluation and Program Planning, 30*, 422–429.

Green, L. (1996). *Policing places with drug problems*. Thousand Oaks: Sage.

Hawken, A. & Kleiman, M. (2009). *Managing drug involved probationers with swift and certain sanctions*. Final report submitted to the National Institute of Justice. Unpublished report.

Hope, T. (1994). Problem-oriented policing and drug market locations: three case studies. *Crime Prevention Studies, 2*, 5–32.

Kennedy, D. (2008). *Deterrence and crime prevention*. New York: Routledge.

Lanza-Kaduce, L., Parker, K. F., & Thomas, C. W. (2000). The devil is in the details: the case against the case study of private prisons, criminological research, and conflict of interest. *Crime & Delinquency, 46*, 92–136.

Lewin, K. (1946). Action research and minority problems. *Journal of Social Issues, 2*, 34–46.

Lipsey, M. W. (1995). What do we learn from 400 research studies on the effectiveness of treatment with juvenile delinquents? In J. McGuire (Ed.), *What works: Reducing reoffending* (pp. 63–78). New York: Wiley.

Lipsey, M. W. (2000). Statistical conclusion validity for intervention research: A significant ($p<.05$) problem. In L. Bickman (Ed.), *Validity and social experimentation: Donald Campbell's legacy* (pp. 101–120). Thousand Oaks: Sage.

Lipsey, M. W. (2003). Those confounded moderators in meta-analysis: good, bad, and ugly. *The Annals of the American Academy of Political and Social Science, 587*, 69–81.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks: Sage.

Mazerolle, L., Price, J., & Roehl, J. (2000). Civil remedies and drug control: a randomized field trial in Oakland, California. *Evaluation Review, 24*, 212–241.

Mazerolle, L., Soole, D. W., & Rombouts, S. (2007). Street-level drug law enforcement: a meta-analytic review. *Campbell Collaboration*. doi:10.4073/csr.2007:2.

Moore, M. H., Prothrow-Stith, D., Guyer, B., & Spivak, H. (1994). Violence and intentional injuries: Criminal justice and public health perspectives on an urgent national problem. In A. J. Reiss & J. Roth (Eds.), *Consequences and control* (Vol. 4, pp. 167–216). Washington, DC: National Academy Press.

National Research Council. (2008). *Parole, desistence from crime, and community integration*. Washington, DC: National Academies Press.

Olds, D. L. (2009). In support of disciplined passion. *Journal of Experimental Criminology, 5*, 201–214.

Petersilia, J. (2008). Influencing public policy: an embedded criminologist reflects on California prison reform. *Journal of Experimental Criminology, 4*, 335–356.

Petrosino, A., & Soydan, H. (2005). The impact of program developers as evaluators on criminal recidivism: results from meta-analyses of experimental and quasi-experimental research. *Journal of Experimental Criminology, 1*, 435–450.

Reason, P., & Bradbury, H. (2008). *The handbook of action research* (2nd ed.). London: Sage.

Rothstein, H. R. (2008). Publication bias as a threat to the validity of meta-analytic results. *Journal of Experimental Criminology, 4*, 61–81.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Sherman, L. W., & Rogan, D. (1995b). Deterrent effects of police raids on crack houses: a randomized controlled experiment. *Justice Quarterly, 12*, 755–82.

Sherman, L. W., & Strang, H. (2009). Testing for analysts' bias in crime prevention experiments: can we accept Eisner's one-tailed test? *Journal of Experimental Criminology, 5*, 185–200.

Sherman, L. W., Buerger, M., & Gartin, P. (1989). *Beyond dial-a-cop: A randomized test of Repeat Call Policing (RECAP)*. Washington, DC: Crime Control Institute.

Skogan, W. G., & Frydl, K. (Eds.). (2004). *Fairness and effectiveness in policing: The evidence*. Washington, DC: The National Academies Press.

Spector, P. E. (1992). *Summated rating scale construction*. Newbury Park: Sage.

Sviridoff, M., Sadd, S., Curtis, R., & Grinc, R. (1992). *The neighborhood effects of street-level drug enforcement: Tactical Narcotics Teams in New York*. New York: Vera Institute of Justice.

Weisburd, D., & Green, L. (1995). Policing drug hot spots: the Jersey City DMA experiment. *Justice Quarterly, 12*, 711–736.

Weisburd, D., Lum, C. M., & Petrosino, A. (2001). Does research design affect study outcomes in criminal justice? *The Annals of the American Academy of Political and Social Science, 578*, 50–70.

Weisburd, D., Telep, C. W., Hinkle, J. C., & Eck, J. E. (2008). The effects of problem-oriented policing on crime and disorder. *Campbell Collaboration*. doi:10.4073/csr.2008.14.

Welsh, B. C., Peel, M. E., Farrington, D. P., Elffers, H., & Braga, A. A. (2011). Research design influence on study outcomes in crime and justice: a partial replication with public area surveillance. *Journal of Experimental Criminology, 7*, 183–198.

Wilson, D. B. (2009). Missing a critical piece of the pie: simple document search strategies inadequate for systematic reviews. *Journal of Experimental Criminology, 5*, 429–440.

Witt, M. D., & Gostin, L. O. (1994). Conflict of interest dilemmas in biomedical research. *Journal of the American Medical Association, 271*, 547–551.

Brandon C. Welsh Ph.D., is a Professor in the School of Criminology and Criminal Justice at Northeastern University and a Senior Research Fellow at the Netherlands Institute for the Study of Crime and Law Enforcement. His research interests include the prevention of delinquency and crime and evidence-based crime policy. He is an author or editor of ten books, including *Experimental Criminology: Prospects for Advancing Science and Public Policy* (Cambridge University Press, forthcoming), *The Oxford Handbook of Crime Prevention* (Oxford University Press, 2012), and *The Future of Criminology* (Oxford University Press, 2012).

Anthony A. Braga Ph.D., is a Professor in the School of Criminal Justice at Rutgers University and a Senior Research Fellow in the Program in Criminal Justice Policy and Management at Harvard University. Dr. Braga's research involves collaborating with criminal justice, social service, and community-based organizations to address illegal access to firearms, reduce gang and group-involved violence, and control crime hot spots. His most recent book is *Policing Problem Places: Crime Hot Spots and Effective Prevention* (Oxford University Press, 2010).

Meghan E. Hollis-Peel M.Sc., is a Doctoral Candidate in the School of Criminology and Criminal Justice at Northeastern University and a Research Associate at the Netherlands Institute for the Study of Crime and Law Enforcement. Her dissertation is titled "Defining Crime, Social Control, and the Enduring Influence of Neighborhood Context: A Mixed Methods Approach."